Predictability and Surprise in Large Generative Models

Year: 2022 | Citations: 325 | Authors: Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova Dassarma, T. Henighan

Abstract

Large-scale pre-training has recently emerged as a technique for creating capable, general-purpose, generative models such as GPT-3, Megatron-Turing NLG, Gopher, and many others. In this paper, we highlight a counterintuitive property of such models and discuss the policy implications of this property. Namely, these generative models have a paradoxical combination of predictable loss on a broad training distribution (as embodied in their ”scaling laws”), and unpredictable specific capabilities, inputs, and outputs. We believe that the high-level predictability and appearance of useful capabilities drives rapid development of such models, while the unpredictable qualities make it difficult to anticipate the consequences of model deployment. We go through examples of how this combination can lead to socially harmful behavior with examples from the literature and real world observations, and we also perform two novel experiments to illustrate our point about harms from unpredictability. Furthermore, we analyze how these conflicting properties combine to give model developers various motivations for deploying these models, and challenges that can hinder deployment. We conclude with a list of possible interventions the AI community may take to increase the chance of these models having a beneficial impact. We intend for this paper to be useful to policymakers who want to understand and regulate AI systems, technologists who care about the potential policy impact of their work, funders who want to support work addressing these challenges, and academics who want to analyze, critique, and potentially develop large generative models.